DOCUMENT RESUME

ED 233 570 FL 013 867

AUTHOR McConochie, Jean Alice
TITLE Simplicity and Complexity in Scientific Writing: A Computer Study of Engineering Textbooks.
PUB DATE 75
NOTE 23p.; Paper presented at the International Congress of Applied Linguistics (1975).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Engineering; English; *English for Special Purposes; Higher Education; *Technical Writing; *Textbook Content

ABSTRACT

The relative frequency and sentence level use of syntactic structures in comtemporary written scientific English was investigated. Two thousand sentences from 100 engineering textbooks were compared with sentences from American literary prose by means of a computer program. The results confirmed the hypothesis that engineering writing uses a smaller subset of the total inventory of grammatical constructions in English than does literary writing. Engineering writers make limited use of the grammatical options available to them, especially in the variety of sector-mix arrays, range and diversity of T-unit length measured in words, number of filled sectors per T-unit, and use of inserts. The typical engineering textbook sentence comprises one T-unit 18 words long with a subject, passive verb, and complement. Engineering writing is a simpler version of English than literary writing. Thus, engineering writing is probably easier for the foreign student of English to learn. (Author/RW)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED233570

# SIMPLICITY AND COMPLEXITY IN SCIENTIFIC WRITING

## A COMPUTER STUDY OF ENGINEERING TEXTBOOKS

Jean McConochie

Center for Applied Linguistics

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jean McConochie

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FL013867

The purpose of the study reported here was to determine the sentence-level syntax of contemporary, written Scientific English. The work grew out of an interest in the reading problems faced by students of English as a foreign language.

The hypothesis of the study was that engineering writing makes use of a smaller subset of the total inventory of the grammatical structures of the language than does literary writing, with the implication that should that be true, then the teaching of a reading knowledge of English as a foreign language to engineers might usefully be limited to that subset.

The primary data for the study consisted of a stratified random sample of 2,000 sentences (approximately 46,000 words) from 100 widely used textbooks in civil and electrical engineering, fields which have accounted for the largest percentage of foreign students in the U. S. in recent years. A Control sample half that size--approximately 23,000 words--was chosen from anthologies of essays used in freshman English courses in American universities.

The availability of a high-speed computer to process the data made it possible to go further than any previous investigation in specifying the grammatical characteristics of scientific writing. The final run of the 15 programs written for the study took two hours of computer time; the total computer time used was close to 15 hours. A similar analysis done by hand--inevitably with less accuracy--would have required some thirty person-years.

Some of the results of all that machine counting are summarized in this paper and the accompanying table. Scientific prose, as represented in the samples, did indeed differ from literary prose in a number of measurable characteristics. But that the hypothesis was supported by the findings is of less interest than the specific points of divergence, for scientific English proved to be both simpler and more complex than literary English prose.

The samples differed, first, in average sentence length, where "sentence" is defined as the words between two full stops. But numerous studies have already suggested the problems of defining a sentence, and a major study (Kučera and Francis, 1967) suggests that sentence-length is, in any case, genre specific.

A much more interesting unit of measure is the T-unit, a term coined by Kellogg Hunt (1965) in his study of what constitutes maturity in writing. T-unit stands for "minimum terminable unit"--an independent clause with its adverbials and linkers.

There is a significant difference in mean T-unit length between literary prose and that produced by writers of engineering texts, that for engineering prose being higher, as shown in section 1.1.1 of the table. This is already a sign of complexity in scientific writing.

Then we may follow Hunt in grouping T-units by length: short (1-8 words), middle-length (9-20), and long (21 plus). Hunt found the increased use of long T-units to be characteristic of the style of mature writers. Nonetheless, in the present study the Control sample has nearly twice as many short T-units as does civil engineering (CE) or electrical engineering (EE). So a preference for middle-length and long T-units over short ones would seem to be a further aspect of complexity in scientific writing.

This tendency is further borne out in modal T-unit length, which is shown in section 1.1.2. The modal range for the Control sample is five words lower than for the two engineering samples.

Grouping T-units by individual lengths, on the other hand, reveals an element of simplicity in scientific writing, for there is much greater variation in T-unit length in the literary sample, with Control predominating in the two lowest ranges of the scale (1-5 and 6-10 words) and in the highest (46 plus). Therefore, wide range and variety of T-unit length characterize literary

but not scientific writing.

Sentences may, of course, contain more than one T-unit; thus there are both simple and compound T-units. (The latter are traditionally called compound sentences). The civil and electrical engineering texts are almost identical in the distribution of simple and compound T-units, as shown in 1.2.1. A majority of the sentences in all three samples have only one T-unit; but a much larger proportion of those in the Control sample are compounds, and only in the Control sample were there sentences composed of four or more T-units. (The immediately preceeding sentence is composed of three T-units.) Taking the number of T-units per sentence as a gross index, the engineering texts show greater simplicity in sentence structure than do literary texts.

Writers of engineering prose characteristically use linking words to join T-units (see 1.2.2), while literary writers are twice as likely to use punctuation alone. The preference of scientific writers for the coordinating conjunction in compound sentences may be traced to the greater notional explicitness of such links.

Minor T-units, as noted in section 1.3, are those which begin with Let or a wh- word, or which are elliptical, for example: "Let us begin." They occur twice as frequently in the Control texts. And even when editorial sentences--those contained in question and discussion sections--are eliminated from the Control sample, the remaining number of minor T-units is still

greater than the total number of minor T-units, including problems at the ends of chapters, in the engineering sample.

From this, we may draw two tentative conclusions: (1) the use of minor patterns for stylistic variation is quite limited in literary writing, accounting for less than three per cent of the T-units, and (2) in scientific writing, this stylistic option is almost never used.

The data were also examined, by T-unit, for sentence level structure, using Sector Analysis (Allen, 1964), a tagmemic grammar which posits 20 positions--or sectors--on the sentence level. A matrix of this size proved unwieldy and was reduced by half to one of 10 sectors: Linker, Front (sentence) adverbial, Subject, X-word (BE or the first auxiliary verb), Middle adverb, Verb, Object, Complement, End (sentence) adverbial, and Postponed subject.

For this 10-sector matrix, the number of filled sectors per major T-unit has a theoretical range of two (S and X or V) to 10 (all sectors filled). The engineering samples are in close agreement both in mean number of sectors used and in variance in relation to the mean. Section 2.1 shows the means for the three samples, that for the Control being higher. The odds are two million to one against the possibility that the difference between the engineering texts and the Control occurred by chance. So in filled sectors per T-unit, we have another aspect

of the relative simplicity of scientific writing.

In number of words per filled sector, CE and EE do not differ significantly. But then, neither do EE and Control. Moreover, the averaging hides long sector-fillers, such as the one cited in 2.2.1, where the C position is filled by a 41-word prepositional phrase.

Aside from the 10 sectors in a T-unit, there may also be "inserts"--constructions which are inserted between sectors and which may be dropped without changing the meaning of a sentence. Insert use further illustrates the relative simplicity of scientific prose. As section 3.1 shows, inserts are used more than twice as often in literary writing as in engineering texts, and there are also more inserts per sentence in the Control sample.

What function do these inserts serve? Explaining them as a means of qualifying a main statement does not account for their predominance in literary texts, for scientific statements are complex in part because of the extensive qualification used in the interests of precise description. The use of inserts in literature seems to be, rather, for humor or irony, as in the sentence from James Thurber (the first example in 3.2), which contains four inserts. In scientific writing, inserts are used for simple reference to diagrams and charts accompanying a text or to achieve greater notional explicitness, as in the second citation in section 3.2.

Let us now consider the ways in which a matrix may

be realized. By definition, literary writers are those most skilled in exploiting the resources of language. In English, their skill is evidenced in part by the use of potential sectors within the sentence, in part by the constructions used within the sectors.

A comparison of the 10-sector matrices revealed that mature writers of English use such a variety of sentence structures that no general patterns emerge. However, a seven-sector matrix proved to be a more sensitive discriminator. This matrix included the F, S, M, O, $C_1$, $C_2$ and E positions. A verbal element was assumed, but whether it was X, V, or X plus V was not indicated.

Sections 4.2.1 and 4.2.2 show the per cent of matrix arrays in the first two ranks, which account for better than one-fifth of all the engineering T-units, but only one-tenth of Control. The most-used pattern accounts for nearly 15 per cent of the T-units in both CE and EE, but only five per cent of the Control T-units. The second- and third-ranked engineering patterns occur with almost identical frequencies, both well below the first rank, while in the Control sample, ranks 1 and 2 have very similar frequencies and the third rank is not far below. Unique matrix arrays (see 4.3) account for some 15 per cent of the engineering T-units and for nearly 20 per cent of Control.

The two most common patterns in the seven-sector matrix are shown in section 4.4 of the table. The most

frequent for all three samples is a cluster (a head-word and its modifiers) filling the subject sector and a prepositional phrase or cluster in the $C_1$ position. The literary control authors use the two C variants--phrase or cluster--with almost equal frequency. The engineering authors, on the other hand, use a prepositional phrase twice as often. It seems possible that this difference is tied to verb type, which was not specified.

This variety of matrix realization illustrates the great diversity of potential patterns in English sentences. The Control authors, as hypothesized, exploit the grammatical possibilities of the language while the engineering writers make extensive use of a highly restricted number of sentence patterns.

Six construction types (ennumerated in section 5) were tabulated for frequency of occurence within the seven-sector matrix. While the rank order of use is the same for all three of the samples, the actual frequencies of use differ considerably.

Clusters and prepositional phrases are more frequent on the sentence level in scientific writing than in literary prose, this being balanced by the greater single-word use in literary writing. CE and Control are similar in sentence-level clausid use; EE is slightly higher. CE has the most included clauses; Control has the fewest. Predicatids are used more often in the Control sample than in the engineering texts.

From these distributions, we can see an important aspect of complexity in scientific prose--a preference for complex constructions (clusters, phrases and included clauses) over single words. An aspect of simplicity is the lesser use of predicatids than is found in literary writing. This reflects, of course, the nominal style of scientific writing as opposed to the more verbal style of literary prose.

Various sector combinations were also tabulated. (As an example, a few of the figures for F, M and E combinations are given in section 6.1.) The clearest distinction in sector combinations comes, as you might expect, in the kind of verb. A combination of X plus a past participle is the most common pattern in CE and is second in frequency for EE, while it is by far the least common pattern in Control. Whether this X-V is a passive or stative verb--a distinction emphasized by Lackstrom, Selinker and Tmimble (1970)--was, unfortunately, not coded. A compound active tense (X plus "non -ed" V) is the least common pattern in both of the engineering samples and ranks third in the Control sample. This confirms Barber's findings (1962) for scientific prose. Literary writers show a strong preference for simple active verbs, which are also the most common verbal type in EE. Examples are given in section 6.2.1

The homogeneity of the findings for civil and electrical engineering texts confirms the assumed existence of an "engineering subset" of English. A typical engineering textbook sentence (what might be called the modal sentence) comprises one T-unit eighteen words long with four sectors filled, giving an average of slightly over four words per sector. The pattern is subject-verb-complement with no sentence-level adverbial. The S position is filled by a cluster, the C position by a prepositional phrase. The verb is either passive or stative. Here are two examples:

S
The electrical system shown in Fig. 6-15(b)
X V C
is operated by an alternating current sup-
plied by an outside source.

S X V
The way to write engineering prose is shown
C
in these examples, which contain all of the
typical elements.

The ease with which a plausible illustrative sentence may be turned out is itself evidence of the relative grammatical simplicity of the texts. Engineering writers make limited use of the grammatical options open to them, notably in variety of sector-matrix arrays, range and diversity of T-unit length, number

of filled sectors per T-unit, and use of inserts.

The complexity of scientific writing, as exemplified in engineering texts, reflects a genre preference for nominal style, as shown in the proportionately higher use of determiners, prepositions, and nominal and adjectival suffixes--morphemes counted though not reported here.

Nevertheless, engineering writing is, on the whole, a simpler version of English than is literary writing. The latter, although also a subset, is better defined by its diversity than by shared characteristics. Syntactic variations appear as elements of individual style--as in Thurber's use of inserts; no patterns of usage are common enough to permit meaningful generalization. Consequently, it would be accurate to say that literary English approximates the characteristics of the English language.

What are the implications of this for teachers and students of English as a foreign language?

Since engineering writing is less complex than literary writing, it is, in a sense, easier to learn. The experience of American graduate students learning German in order to pass a language examination bears this out. Such students can read a German encyclopedia far more easily than the works of Goethe, for the more literary the language, the more difficulty there is in understanding, interpreting, and appreciating its rhetorical effectiveness.

In dealing with a subset of any language, then, the student's task is essentially that of deciphering a code. In preparing engineering students to read English, the EFL teacher may usefully begin by teaching students to break this code.

Much of the simplicity of scientific writing lies in the use of a few sentence-level patterns. Once a student is able to recognize these patterns, identifying the component tagmemes and understanding their syntactic and logical relationships, he will be better able to deal with the complexity of individual sector fillers. It is the responsibility of the teacher and textbook writer to provide sufficient practice in identifying those sentence patterns most commonly used.

The high percentage of passive/stative verbs used in engineering writing points up the necessity for students to learn the forms of be and of regular and irregular past participles. In addition, engineering students must be able to recognize the structure of included clauses and their use as sentence adverbials and postponed subjects.

While function words account for some forty per cent of the running words in the text, the findings suggest that relatively few different function words are used. If this proves to be general, it will mean that once the technical vocabulary is mastered, the addition of a small number of words can provide great reading power.

Other questions which have been raised by this study suggest possible directions for future research

into both lexical and grammatical matters. For a given genre, do included clauses manifest the same tagmemes as non-included clauses? For example, is the proportion of active and passive/stative verbs the same on the sentence and clause levels? The nature of the lexical constraints between sectors is another area of interest. Which lexical patterns are most frequent in specific genres? What are the differences in frequency distribution of word classes for different genres?

If it should be possible to answer these and other questions for spoken English as well as written, the issue of priority between these two forms in EFL teaching will become one that can be explicitly rather than impressionistically handled.

I hope that this study will stimulate some teachers of English as a foreign or second language to develop their own applications for the findings of grammatical research, and that it will encourage other scholars in the humanities to regard the computer, like the typewriter, as an easily accessible and invaluable tool.

---

Paper given at the Fourth International Congress of Applied Linguistics, Stuttgart, 1975.

COMPARATIVE FEATURES OF CE, EE AND CONTROL SAMPLES

| | Feature | CE | EE | Control | Comments |
|---|---|---|---|---|---|
| 1.0 | T-UNITS | | | | |
| 1.1.1 | Length (no of wds.) | 19.56 | 18.25 | 17.98 | No sig. diff. btwn. CE & EE; CE/Control & EE/Control diff. at better than .01 level. |
| 1.1.2 | Distribution by length in wds. (modal range) | 16-20 16 | 16-20 | 11-15 | Control has more short T-units (2x CE; 1.3x EE) & greater range & variety of lengths. |
| 1.2.1 | % compound T-units | 22.50 | 22.10 | 34.10 | Lit. writing uses more compound T-units (only Control had 4 or more in one sentence). |
| 1.2.2 | % T-units joined by punctuation alone | 17.80 | 12.00 | 32.00 | Fewer coordinating conjunctions in Control. |
| 1.3 | % minor T-units | 2.24 | 2.53 | 5.92 | Minor patterns twice as frequent in Control. |
| 2.0 | SECTORS | | | | |
| 2.1 | Filled sectors/T-unit (min. = 2; max. = 10) | 4.25 | 4.15 | 4.45 | Diff. btwn. means for CE/EE (4.20) & Control approx. 5 Standard Errors. |
| 2.2 | Words/filled sector | 4.76 | 4.55 | 4.32 | CE/Control diff. at better than .01; CE/EE & CE/Control diffs. not sig. at .01. Averaging hides long sector fillers, e.g. 2.2.1. |

2.2.1 Example of long sector-filler:

It may be defined as that point lying on the line connecting the principal point and the plumb which is at a distance from the former equal to the focal length of the photograph multiplied by the tangent of half the angle of tilt. (CE 5:234)

| | Feature | CE | EE | Control | Comments |
|---|---|---|---|---|---|
| 3.0 | INSERTS | | | | |
| 3.1 | % T-units with inserts | 9.5 | 8.2 | 22.1 | Sentence-level inserts twice as frequent in Control. |
| 3.2 | Examples of inserts: | | | | |

$I_1$ $I_2$

I called the other afternoon on my old friend, Graves Moreland, the Anglo-American literary critic--his mother $I_3$ was born in Ohio--who lives alone in a fairy-tale cottage on the Upson Downs, raising hell and peacocks, the former only when the venerable gentleman becomes $I_4$ an angry old man about the state of literature or something else that is dwindling and diminishing, such as human stature, hope, and humor. (Control 7:295; James Thurber)

I

. . . average torque is developed in the direction to increase $O_o$ and thereby to keep the rotor revolving. (EE 16:81)

| | | | | | |
|---|---|---|---|---|---|
| 4.0 | 7-SECTOR MATRIX ARRAY (F, S, M, O, $C_1$, $C_2$, E) | | | | |
| 4.1 | Total no. diff. arrays | 253 | 255 | 307 | Variety 20% greater in Control. |
| 4.2.1 | % T-units in Rank 1 | 14.6 | 13.7 | 5.6 | 1st 2 ranks = 20+% CE/EE, 10% Control. |
| 4.2.2 | % T-units in Rank 2 | 8.6 | 8.2 | 5.3 | After Rank 2, some agrmnt. btwn. CE & EE, |
| 4.3 | % T-units with unique matrix array | 14.4 | 14.7 | 19.6 | almost none btwn. them & Control. |

15.

| | Feature | CE | EE | Control | Comments |
|---|---|---|---|---|---|
| 4.4 | 2 most common arrays: | | | | |
| 4.4.1 | Rank: | 1 | 1 | 2 | |
| | S (cluster) | | C (prep. phrase) | | |
| | The materials are arranged in order of increasing conductivity. | | | | |
| 4.4.2 | Rank: | 2 | 2 | 1 | |
| | S (cluster) | | C (cluster) | | |
| | The exponent in the equation is approximately correct. (CE 43:416) | | | | |

5.0 CONSTRUCTION DISTRIBUTION IN THE 7-SECTOR MATRIX (in rounded %)

| | | | | |
|---|---|---|---|---|
| Cluster | 44 | 41 | 36 | While rank order of use same for all, CE & EE use more cluster & fewer single words. |
| Prepositional phrase | 26 | 26 | 21 | |
| Single word | 14 | 17 | 26 | *A clause with a non-time-oriented verb, e.g. the unit soil weight multiplied by the area of triangle CGH (CE 1:15) |
| Clausid* | 9 | 11 | 9 | |
| Predicatid** | 5 | 4 | 6 | **A non-time-oriented verb + its objects &/or complements, e.g. to specify the orientation of a crystal plane by its Miller indices (EE 49:58) |
| Included clause | -1 | -1 | -1 | |

6.0 USE OF SELECTED SECTORS IN THE 7-SECTOR MATRIX (in rounded %)

| | | | | | |
|---|---|---|---|---|---|
| 6.1 | F, M, E - % of potential positions used: None | 59 | 59 | 56 | Rank order of FME use same for all samples. More than ½ each sample, no T-unit adverbials. |
| | F only | 23 | 20 | 22 | F alone much more common than M or E alone or in combination. |

| | Feature | CE | EE | Control | Comments |
|---|---|---|---|---|---|
| 6.1.2 | Example of F filler: | | | | |
| | Frequently the transformation is done in steps. (EE 73:180) | | | | |
| | If deflections are computed at several places, an average value of $P_{cr}$ will be determined which more nearly approximates the true value. (CE 30:357) | | | | |
| 6.2 | X and V (% distribution) | | | | |
| | XV$_1$ (passive/stative) | 36.4 | 28.9 | 10.1 | Passive/stative most common pattern in CE, least |
| | -V (simple active) | 31.1 | 32.8 | 42.5 | common in Control. Compound active tense least |
| | X- (aux. alone) | 21.7 | 25.7 | 25.4 | common pattern in CE & EE (cf. Barber, who found |
| | XV$_2$ (compound active | 10.8 | 12.4 | 22.0 | same in broader sample of scientific writing). |
| 6.2.1 | Examples of most common verb patterns: | | | | |

XV$_1$: Effective expression, both oral and written, can (X) be developed (V) in debate and school publications. (CE 28:6)

-V: The wave of current through a pure capacitance leads (V) the wave of voltage . . . . (EE 44:37)

X-: . . . but at stellar temperature the agitation is (X) so great that . . . . (CE 12:204)

XV$_2$: Each term must (X) have (V) the same dimensions . . . . (CE 43:156)

## BIBLIOGRAPHY

Allen, Robert L. English Grammars and English Grammar. New York: Columbia University Institute for Teachers of English, 1964.

Barber, C. L. "Some Measurable Characteristics of Modern Scientific Prose." Contributions to English Syntax and Philology. Gothenburg Studies in English, No. 14. Göteborg: Elanders Boktryckeri Aktiebolag, 1962.

Hunt, Kellogg W. Grammatical Structures Written at Three Grade Levels. NCTE Research Report No. 3. Champaign, Ill.: National Council of Teachers of English, 1965.

Lackstrom, John E., Selinker, Larry, and Trimble, Louis P. "Grammar and Technical Writing." English as a Second Language: Current Issues, ed. Robert C. Lugton. Language and the Teacher: A Series in Applied Linguistics, No. 6. Philadelphia: Center for Curriculum Development, 1970.

Kučera, Henry, and Francis, W. Nelson. Computational Analysis of Present-Day American English. Providence, R. I.: Brown University Press, 1967.

McConochie, Jean A. "Simplicity and Complexity in Scientific Writing: a Computer Study of Engineering Textbooks." Unpublished Ed.D. project report, Department of Languages and Literature, Teachers College, Columbia University, 1969.